Bias Estimates in Bootstrapping

Bettonvil, B.W.M.

Publication date:
1999

BIAS ESTIMATES IN BOOTSTRAPPING

Bert Bettonvil
Department of Information Systems
Tilburg University
5000 LE Tilburg, The Netherlands

Abstract

Five bootstrapping techniques are examined and compared on their potential for bias reduction. Four of these techniques are from literature, and one is new. An extensive study on a literature example reveals the superiority of the new technique.

Keywords: bootstrapping, bias reduction.

1. INTRODUCTION

The bootstrapping technique originates from 1979 (Efron, 1979) and it is extensively described in Efron and Tibshirani, 1993 (here abbreviated to E&T). Chapter 10 of the latter book describes two estimates of bias, one we shall call the classic estimate, and one E&T call the improved bias estimate. Both methods, described in sections 2.1 and 2.2 respectively, use *nonsystematic resampling*, that is, B random samples of size n (which is the size of the original sample) are drawn with replacement from that original sample .

Davison *et.al.* (1986, briefly mentioned in E&T) present the idea of systematic resampling in, what they call, "balanced bootstrap simulation". To avoid confusion with the methods to be presented later, we call this method "globally balanced", see section 2.3. Instead of sampling with replacement Davison *et.al.* propose to sample without replacement from the set containing B times the original sample. This may also be described as a random permuta-

tion of B times x_1, ... , B times x_n.

Graham *et.al.*.(1990) combine the use of latin squares with permutations of the indices of the selected observations. The use of permutations only will be considered in this section 2.4 under the name of "local balance".

As a new method, in section 3 a distance measure between samples is introduced. We consider the mean squared distance between all possible bootstrap sample and the original sample, and we propose to choose the bootstrap samples in such a way that the mean squared distance between the bootstrap samples chosen and the original samples is equal to the quantity mentioned above.

Section 4 contains focusses on the E&F (pp.126-133) patch data example. The five bootstrap method are compared in various ways. The classic bootstrap bias estimate turns out to be inferior, whereas, in terms of the mean squared error, the newly presented method is best.

2. OVERVIEW OF PREVIOUS WORK

2.1. The classic bootstrap bias estimate

Bootstrapping is a jackknife-related nonparametric method for assessing properties of the population underlying a given "original" sample x_1, ... , x_n. A classic bootstrap sample consists of n draws with replacement from the original sample. The parameter of interest, θ, is estimated by $\hat{\theta}$, which is some function of the sample, say, $\hat{\theta}=t(x_1, \ldots , x_n)$.

Let us denote a bootstrap sample by $x_1^b,\ldots,x_n^b$, then we can compute the bootstrap estimate θ^b as $\theta^b=t(x_1^b,\ldots,x_n^b)$; the latter is called the plug-in estimate. With B bootstrap samples we can compute B bootstrap estimates $\theta^1,\ldots,\theta^B$.

The E&T *classic bootstrap bias estimate* is defined as the difference of the mean of the B bootstrap estimates (let us call this mean $\tilde{\theta}$) and the sample estimate $\hat{\theta}$: $\hat{\beta}=\tilde{\theta}-\hat{\theta}$. To put it loosely: going from the sample to the bootstrap gives the same bias as going from the population to the sample.

2.2. The improved bootstrap bias estimate

The classic bootstrap bias estimate is sensitive to the *unbalance* in the B bootstrap samples; that is, some of the x_i (i=1,...n) appear less than B times in the bootstrap samples, whereas some appear more than B times. E&T propose to compensate for this effect by constructing the n-vector $\bar{P}$, of which entry i (i=1,,,n) consists of the number of occurrences of observation i in the B bootstraps, divided by nB. They argue that for each bootstrap sample there exists an n-vector P^b, of which entry i consists of the number of occurrences of observation i, divided by n, so that there exists a mapping from $T:\mathbb{R}^n \to \mathbb{R}$ such that $T(P^b)=t(x_1^b,...,x_n^b)$ and, with $P^0=(1/n,...,1/n)$, $T(P^0)=t(x_1, ... , x_n)$. The classic bootstrap bias estimate can now be written as $\hat{\beta} = \Sigma_{b=1}^{B} T(P^b) - T(P^0)$, and E&T define the *improved bootstrap bias estimate* as $\hat{\beta} = \Sigma_{b=1}^{B} T(P^b) - T(\bar{P})$.

E&T implicitly assume that $T(\bar{P})$ is easily computed; that, however, is not always the case. In Gifi (1990) a number of non-linear multivariate techniques is described, under the assumption that each observation appears a discrete number of times. Here $T(\bar{P})$ is computable, but if the original data matrix consists of n rows and m columns, then the computation of $T(\bar{P})$ needs a data matrix of Bn rows and m columns, and the computation time may increase by a factor of B^2.

In my own practice I encountered the next problem with trace-driven simulation; see Kleijnen, Cheng and Bettonvil (forthcoming). Suppose for n days the arrival times and service times for k customers are given, and some statistic like mean waiting time or some percentile is computed. Bootstrapping is next performed in the following way: we consider the n sets of k service times as a sample of size n from a k-dimensional space, and bootstrapping means sampling with replacement from n k-tuples. Here we have the peculiarity that the resampling vectors P^* do not uniquely correspond to bootstrap samples, because the order of sampling is crucial. As a consequence, $T(\bar{P})$ is undefined.

The following methods resample in such a way that $\bar{P} = P^0$.

2.3. Globally balanced bootstrap

Although bootstrapping is defined as sampling with replacement, Davison *et.al.* (1986) propose to sample without replacement from the set consisting of B x_1's, B x_2's, ... , B x_n's; or, equivalently, to take a random permutation of the set consisting of B 1s, B 2's, ... , B n's. This

guarantees that $\bar{P}=P^0$, and so $T(\bar{P})=\hat{\theta}^*$. This approach may seem to cause computational problems, but in fact these problems are minor.

Let, at any moment, c_i denote the number of observations i that has to be drawn; the selection begins with all $c_i=B$; it ends with all $c_i=0$. Let c^m be the minimum of the c_i (i=1,...,n). Let R denote the total number of observations that is left to be drawn (at the beginning R=nB). Let r be a random number from U(0,1). If $rR \leq nc^m$, then select observation $j=rR/c^m$, rounded up. If necessary (that is, if $c_j=c^m$), diminish c^m by 1. If $rR>nc^m$, then reduce $rR-nc^m$ by c_1-c^m, c_2-c^m, et cetera, until the remainder is negative. This gives the observation to select.

This approach guarantees that the number of bootstrap observations is balanced; at the price that the number of equal observations tends to be lower than in E&F's bootstrapping. However, when B is large, this disadvantage will diminish. Our next alternative does not have this disadvantage.

2.4 Locally balanced bootstrap

Graham *et.al.* (1990) introduce a way of selecting bootstrap samples that uses latin squares on the one hand, and permutation of indices on the other. Latin squares limits their approach to powers of integers, which may be considered too rigorous a limitation. Permutation of indices, however, is attractive, so we discuss this now.

Draw a bootstrap sample $\{x_i^*\}_{i=1,..,n}=\{x_{\xi(i)}\}_{i=1,..,n}$ where each $\xi(i)$ is randomly drawn from {1,2,...,n} with replacement, and select a random permutation P of {1,2,...,n}; that is, the function $P:\{1,2,...,n\}\rightarrow\{1,2,...,n\}$ is such that $\{P^k(1)\}_{k=1,....,n}=\{1,2,...,n\}$ and $P^n(1)=1$. Besides the original bootstrap sample $\{x_{\xi(i)}\}_{i=1,..,n}$ we also consider $\{x_{P(\xi(i))}\}_{i=1,..,n}$, $\{x_{P^2\xi(i)}\}_{i=1,...n}$, ..., $\{x_{P^{n-1}\xi(i)}\}_{i=1,...n}$. By repeating this selection B' times, we arrive at B=nB' samples.

Graham *et.al.* only consider the permutation $P(i)=i+1$ $(1\leq i<n)$, $P(n)=1$, whereas we propose to use a new random permutation for each new sample $\{x_{\xi(i)}\}_{i=1,..,n}$ to avoid possible systematic effects.

3. BOOTSTRAP WITH BALANCED DISTANCE

From now on we describe bootstrapping as sampling, not from $\{x_1, ... , x_n\}$, but from {1,...

,n}, which is only a matter of notation. Consider the set of all n-tuples of the numbers 1,...,n. This set contains n^n members. The correspondence between a bootstrap sample and an n-tuple is obvious.

Let m_{it} (i=1,...n, t=1,...,n^n) denote the number of i's in tuple t. Then

(a) the mean of m_{it} over all tuples equals 1;

(b1) the mean of m_{it}^2 over all tuples equals (2n-1)/n; and

(b2) the mean of $m_{it}m_{jt}$ over all tuples, equals (n-1)/n; for all i and j (i≠j).

The fact that (a) the mean of m_{it} over all tuples, equals 1, for all i, follows immediately from symmetry. Proposition (b1) is implied by the second moment about zero of the binomial distribution, which is $np+n(n-1)p^2$ see, e.g. Johnson, Kotz, and Kemp (1992), with p=1/n. Proposition (b2) is implied by the mean and covariance of the multinomial distribution, which are np and $-np^2$, respectively (see, e.g. Evens, Hastings and Peacock, 1993), which reduce to 1 and -1/n by substitution of p=1/n.

We now can split up the set of all n-tuples into n^{n-1} subsets of size n in such a way that - within any subset - the mean of m_{it} over all tuples equals 1 (property a). One possible way of doing this is by permutation, which we call local balancing.

Next we want to split up the set of all n-tuples into 2^{n-2} subsets of size n^2 in such a way that - within each subset - not only the mean of m_{it} over all tuples equals 1, but also the mean of m_{it}^2 equals (2n-1)/n, and the mean of $m_{it}m_{jt}$ over all tuples, equals (n-1)/n; for all i and j (i unequal to j). If this were possible, this would give us a way to identify the subsets. Unfortunately, this is not possible in general; it is only possible if n is a power of a prime (private communication with dr. Henny Wilbrink of Eindhoven University of Technology). Because it is a natural generalization of local balance (based on property (a)) to try to implement properties (b), we thought it useful to mention this impossibility, to prevent other researchers from trying to extend local balance in this direction. We switch to another approach.

We map the n-tuples of the numbers 1,2,...n onto the n-tuples consisting of the numbers m_i; that is, if the original tuple contains m_i times the digit i, then entry number i of its mapping is m_i. The latter n-tuples can be viewed as points in an (n-1)-dimensional space, with (1,1,...,1)' being the centre of all tuples. We can define the distance d between any tuple and

the centre as the square root of $\sum_i (m_i-1)^2$.

The mean of d^2 over all possible tuples equals n-1, as can be seen as follows. For i fixed, the mean of $(m_i-1)^2$ is $Em_i^2-2Em_i+1$. The first term equals (2n-1)/n, as we saw before; and $Em_i=1$; so the mean of $(m_i-1)^2$ equals (2n-1)/n-2+1 = (n-1)/n. Summation over all i=1,2,...,n gives the desired result.

Therefore we propose to choose the observations such that the demand of local balance is satisfied, and moreover, that the mean of the squared distances equals n-1. We call this method "bootstrap with balanced distance".

The method is implemented as follows. The bootstrap sample size B is chosen as a multiple of n. Then B/n times a bootstrap sample is drawn, plus a random permutation of {1,2,...,n}. The bootstrap sample indexes are permutated n-1 times, and the squared distance is computed. For the last sample, if the mean of the squared distances is not equal to n-1, then a new sample is drawn. If this fails n times (this number of times is rather arbitrary), then the whole procedure is repeated. In the example presented in the next section, this procedure works very well (because the possible squared distances are 0,2,...,26, and 30,32,42, and 56). Still, we are looking for a better way to construct bootstrap samples.

4. COMPARISON OF THE FIVE BOOTSTRAP METHODS IN AN EXAMPLE

Example. We use an example due to E&F, p.127-133. We have eight observations, given in table 1, and the parameter of interest is

$$\theta = \frac{E(\text{newpatch}) - E(\text{oldpatch})}{E(\text{oldpatch}) - E(\text{placebo})} = \frac{Ey}{Ez}.$$

The estimate of θ is $\hat{\theta} = \frac{\bar{y}}{\bar{z}}$, which is -0.0713.

Next we bootstrap the data in our five ways, using 16, 32, 48, 96, 192, 400, 800, 1600 and 3200 bootstraps; E&T used 25,50,...,3200 bootstraps, but two of our methods require a multiple of eight bootstraps. For every number of bootstraps every method is replicated 20 times (E&T do not replicate, which makes their results less reliable).

Figure 1 gives the second largest and the second smallest bias estimate for all numbers

of bootstraps and for all methods. From figure 1 we learn that classic bootstrapping is inferior to the other methods, which are all about equally good. Removing classic bootstrapping gives us figure 2 (with a different y-scale). Methods two, three, and four all give the highest upper bound at some instances, and the lowest lower bound at some other instances. This indicates that the method with balanced distance is to preferred.

TABLE 1.

The patch data according to Efron and Tibshirani, p.127

subject	placebo	oldpatch	newpatch	z=old-plac	y=new-old
1	9243	17649	16449	8406	-1200
2	9671	12013	14614	2342	2601
3	11792	19970	17274	8187	-2705
4	13357	21816	23798	8459	1982
5	9055	13850	12560	4795	-1290
6	6290	9806	10157	3516	351
7	12412	17208	16570	4796	-638
8	18806	29044	26325	10238	-2719
mean:				6342	-452.3

As an indication of the spread, figure 3 gives the mean of the fifth and sixth largest and smallest bias estimate. Again, classic bootstrap is the worst method. Removing this method gives figure 4. Using the criterion “which is ever the worst” again favours method five.

Figure 5 gives the mean squared error: the conclusion is evident and is the same as mentioned before. In figure 6 classic bootstrap is removed, and again method five is never the worst.

Finally, figure 7 gives log mse, and demonstrates the superiority of the method of balanced distances.

Note that the use of mse and log mse assumes the expectation of the bias to be known.

E&T claim the expectation is known to be .0079, which is computed by the classic method with 100,000 draws. We repeated this 20 times and found a mean of .007672 and a standard deviation of .000306. So E&T's estimate is less than one s.d. removed our mean. We also computed the bias for 1,000,000 draws, which gave a mean of .0077267 with s.d. .000099.

However, using the improved bias estimate for 100,000 draws gives a mean of .007761 and an s.d. of .000056; for 1.000,000 draws we found the same mean and an s.d. of .000018. Because that the improved method gives more stable results - and E&T regard this method better than the classic method - we used .007761 as the mean in the computation of the mean squared error.

5. CONCLUSION AND FUTURE RESEARCH

Based on an example of Efron and Tibshirani (1993) two conclusions emerge. First, the classic bootstrap bias estimate is by far the least favourable. Second, bootstrap with balanced distance outperforms its competitors, which perform roughly the same.

Two issues are left for future research. First, the construction of a balanced distance bootstrap sample needs reconsidering; second, comparisons may be made using other bootstrap applications.

Acknowledgement. The author wishes to thank dr. Henny Wilbrink of Eindhoven University of Technology for his advice on Galois Fields, and Dr. Jack Kleijnen for his comments on an earlier version of this paper.

LITERATURE

Efron, B. (1979), Bootstrap methods: another look at the jackknife, *Ann. Statist.***7**, 1-26.

Efron, B., and R.J. Tibshirani (1993), *An Introduction to the Bootstrap*, Chapman & Hall.

Evans, M., N.Hastings, and B.. Peacock (1993), *Statistical Distributions*, Wiley.

Davison, A.C., D.V. Hinkley, and E. Schlechtman (1986), Efficient bootstrap simulation, *Biometrika* **73**, 555-66.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, Wiley.

Graham, R.L., D.V. Hinkley, P.W.M. Johen, and S. Shi (1990), Balanced Design of Bootstrap Simulations, *J.R.Statist.Soc.B* **52**, 185-202.

Johnson, N.L., S.Kotz, and A.W.Kemp (1992), *Univariate Discrete Distributions*, Wiley.

Kleijnen, J.P.C., RC.H.Cheng, and B.Bettonvil (forthcoming), Validation of trace-driven simulation models: Bootstrap tests.